OBJECTIVE EVALUATION OF WORD PRONUNCIATION BY FILTER-BAND ANALYSIS

Mark J. Bakkum, Reinier Plomp and Louis C.W. Pols*

Department of Otolaryngology, Free University Hospital
P.O.box 7057, 1007 MB AMSTERDAM, The Netherlands
*Institute of Phonetic Sciences, University of Amsterdam

ABSTRACT

Usually, pronunciation is evaluated subjectively by listening. The aim of this research project is to obtain an objective measurement of the quality of pronunciation. For this objective evaluation a real-time spectral analyzing method is developed on a digital signal-processor (16 bandfilters according to the critical-band model). For every word this method leads to a different trace in a 16-dimensional spectral space. Besides level and speaker normalisation also time normalisation will be applied by determining an average spectrum for each phoneme and transition.

CVC-words were analyzed spoken according to the Dutch pronunciation rules by male adults (6 native Dutch and 9 foreigners). The feasibility of the objective evaluation has been investigated by considering whether the spectral information as expressed in various distance measures gives an adequate description of subjective judgements of the phonemes.

INTRODUCTION

The average listener has, if not hearing impaired, no problems in recognizing words spoken in his own language by different speakers. He is also able to judge the quality of the pronunciation in general terms, which enables him to determine whether the speaker spoke in his native language or not, whether the speaker has an accent and whether he spoke with obvious abnormalities. These judgements, however, are subjective, which implies that they will vary for different listeners and that, for one listener, they will vary in time. In order to investigate the processes involved in the perception of speech and to provide a method for objective evaluation and recognition a simulation of hearing by a machine would be very useful.

Speech is characterized by properties as duration, intensity, fundamental frequency and spectrum. Previous research (cf. Pols [7], Assmann and Summerfield [1]) shows that the spectral properties give most information. As our approach is perceptive-oriented rather than production-oriented it seems reasonable to apply a bandfilter analysis in which the bandfilters correspond with the ear's critical bandwidth. This method, introduced by Plomp et al. [6], has been adopted by several other investigators. Carlson and Granstrom [2] and Klatt [3] proved, using synthetic vowels, that phonetic distance could be described this way and Pols et al. [8] reported good agreement between spectral and perceptive data for natural speech. Also recognition tasks were performed quite well (Pols [7]).

We investigated whether the method may be applied to judge articulatory pronunciation differences on phoneme base for various speakers. By segmenting in phonemes with more or less constant spectra we emphasise the static aspects of speech. The possibly important role of the transitions was not yet considered. Because of the substantial spectral differences between males and females at first only speech of males has been analyzed.

A REAL-TIME DIGITAL FILTERBAND-ANALYSIS SYSTEM

In order to obtain spectral representations of spoken words a real-time digital filterband analysis system has been designed. The system consists of an Audio Signal Processor, based on the TMS 32010 processor chip, the necessary devices for input and output and a PDP 11/23 or a PC/AT host computer for control and communication. All possible digital filter designs can be implemented on this system.

We have chosen a set of 16 passband filters for which the transfer characteristics are given in Fig. 1. These are recursive, third-order, elliptic filters, realized by a cascade of three second-order 'direct form II' sections (see Tretter [9]). The frequency range covered by the

![Fig. 1. Attenuation characteristics of the set of 16 passband filters.](image-url)
filters is from 90 to 7200 Hz and the bandwidth of the filters is 1/3 octave, except for the lowest three filters, for which the bandwidths are, respectively, 1, 0.6 and 0.4 octave. These bandwidths are in good agreement with the critical bandwiths.

A scheme of the complete system is given in Fig. 2. To make real-time calculations possible, three TMS 32010 processors operate in a parallel way. Communication between these processors takes place through their collective memory. Communication on the base of vector interrupts is possible between the host computer and processor 1. Precise timing is regulated internally.

To spare valuable time, the input signal is sampled at four different frequencies, so four 12-bit A/D-convertors and corresponding analogue low-pass filters are used. For the three highest 1/3 octave filters (3600-7200 Hz), the sampling frequency was 20 kHz. For the three filters that cover the octave below (1800-3600 Hz) the sampling frequency was 10 kHz, so that not only half of the calculation time is needed but also exactly the same filter designs could be used. The octaves below were sampled with, respectively, 5 and 2.5 kHz, so that the intensities for input and all 16 filters could be calculated every 12.8 ms. The spectra of the spoken input are thus immediately available between the host computer and processor 1. Precise timing is regulated internally.

Recordings were made using a high-quality microphone, followed by an amplifier and a high-pass-filter (cut-off frequency 100 Hz, 12 dB/octave) to eliminate ambient low-frequency noise.

Fig. 2. Scheme of the filterband analysis system.

OBJECTIVE, PHYSICAL ANALYSIS OF SPOKEN WORDS

1. Acquisition of level and speaker normalized spectra of phonemesegments

Recordings of lists of words were made for 15 male speakers in a well-insulated soundproof room. Six of them were native Dutch, in the age of 28 to 45 years, speaking without strong accents or obvious abnormalities. They were asked to read out 12 typed lists of 30 words at a steady rate at a normal level. The other nine speakers were foreign students, six Moroccans, one Egyptian, one Indonesian and one Portuguese. They all followed a Dutch course for beginners, their ages varied from 19 to 31 years and their stay in Holland varied from 3 to 26 months. The words were presented to them in sets of three, visually on a terminal screen as well as auditivey and they were asked to repeat these words at about the same rate at a normal level. As examples words were used spoken by one of the native Dutch, a well-trained speaker without any accent. When obvious errors or disturbances occurred, the speaker was asked to repeat that particular set. In order to preserve their concentration, speakers could take some breaks.

The list consisted of 360 CVC words, among which the combinations of all initial and final consonants possible in Dutch language, with, respectively, final /t/ and initial /h/, for each of the vowels /a/, /i/ and /u/. Furthermore, all 15 different Dutch vowels and diphthongs were presented three times in the most neutral h-t context.

To be able to compare the spectral information of different utterances of one word we have to eliminate variations in length. The influence of these variations on speech quality should be examined separately. For this reason we segmented the words into the more or less steady parts of the phonemes and the transitions. This segmentation was carried out by a program that makes use of the level differences to isolate the vowel and of spectral differences and a priori knowledge to detect the consonants and transitions. Because of the difficulties in segmentation caused by the large differences among speakers results were always checked by listening and by considering the oscillogram and spectrogram.

At first we examined the evaluation method when applied to the 15 possible vowels and diphthongs. This means that we put most emphasis on the static aspects of speech. The dynamical aspects of the transitions, which are probably harder to judge but also important for speech quality, are left out of consideration at this stage. For all speakers one /h(vowel)t/-utterance was selected for all 15 vowels. Criteria for the selection were quality and level of the recordings and, most important, the duration of the vowel segment. These durations should be comparable, because in the corresponding subjective experiment we want to present the subjects vowel segments of equal duration so that only spectral, articulatory properties are judged. Because of their steadiness, for most vowels no problems arise when only part of it belongs to the segment. We had to be more careful for the diphthongs but a satisfactory selection proved to be possible. Durations of the segments are given in table 1.

### Table 1. symbols and segment durations of vowels and diphthongs

<table>
<thead>
<tr>
<th>phonetic</th>
<th>a</th>
<th>a</th>
<th>c</th>
<th>i</th>
<th>e</th>
<th>i</th>
<th>e</th>
<th>o</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>orthographic</td>
<td>aa</td>
<td>a</td>
<td>e</td>
<td>i</td>
<td>ee</td>
<td>ee</td>
<td>i</td>
<td>e</td>
<td>oo</td>
</tr>
<tr>
<td>duration (ms)</td>
<td>115</td>
<td>77</td>
<td>64</td>
<td>64</td>
<td>115</td>
<td>64</td>
<td>128</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>phonetic</td>
<td>u</td>
<td>a</td>
<td>y</td>
<td>ø</td>
<td>au</td>
<td>Ay</td>
<td>el</td>
<td></td>
<td></td>
</tr>
<tr>
<td>orthographic</td>
<td>oe</td>
<td>u</td>
<td>uu</td>
<td>eu</td>
<td>au</td>
<td>ui</td>
<td>e</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
For the selected segments the average transfer spectrum was determined. We eliminated the effects of level differences by equalizing their average levels. We also applied speaker normalization to compensate for the spectral effects of unchanging non-articulatory speaker-specific characteristics such as the sizes of the vocal tract and shape and functioning of the sound-production source (Mulleenix et al. [4] showed that adaptation to speaker properties plays a role at perception). The average of all vowel spectra of one speaker was calculated and this average spectrum was then subtracted from the individual spectra.

2. Distance measures for comparison of spectra

In many ways distances may be calculated between the transfer spectra of the vowel segments. First of all it seems to be reasonable to search for a standard spectrum that represents a vowel of undisputed good quality. Assuming at first approach that such a standard is more or less unique, we determined the standard for each vowel as the average of the normalized spectra of that particular vowel as spoken by all six native Dutch speakers. Only if one of these spectra showed a relatively great deviation from the others it was excluded. It will be clear that in this way only great differences in vowel quality can be explained by the evaluation method. However, listening showed that these great differences are surely present in the available vowel data as spoken by the nine foreigners.

A distance measure as suggested by Plomp [5] is the Euclidian distance between two spectra in the 16-dimensional spectral space. When two spectra N and X are defined by their DB levels Lni and Lxi this measure DP (in dB) is defined as follows

$$DP_{nx} = \sqrt{\sum_{i=1}^{16} L_{ni}^2 - L_{xi}^2}.$$  

This measure is in agreement with the apparently logarithmic processing of level differences by the human ear. Many experiments have shown the usefulness of this measure, but some researchers tend to use a derived measure that is based on formant-peak finding mechanisms (Klatt [3] and Assmann & Summerfield [1]). Here, differences in spectral slopes form the base of the measure while these differences are weighted more heavily when occurring in formant peaks. The 'Klatt-metric' DK (in dB) is computed as follows

$$DK_{nx} = \sum_{i=1}^{16} 0.5 (kG + kL)(S_{ni} - S_{xi})^2,$$  

where $S_{ni}$ and $S_{xi}$ are the spectral slopes

$$S_i = L_i - L_{i+1},$$  

for $i = 1$ to $15$, while $S_{16} = S_{15}$. The weighting function k is defined by

$$k_i = \frac{L_i}{(D_{G1} + D_{L1})} \times \frac{L_i}{(D_{G1} + D_{L1})},$$  

where $D_{G1}$ is the difference between $L_1$ and the global maximum and $D_{L1}$ is the difference between $L_1$ and the nearest local maximum. The smaller the values assigned to $kG$ and $kL$ are, the more sensitive the measure will be to differences near maxima. Values may be assigned such to give well agreement with possible subjective data. Klatt [3] suggested values of 20 for $kG$ and 1 for $kL$.

RESULTS AND DISCUSSION

The average Euclidian spectral distances to the standard for all 15 vowels and diphthongs were plotted in fig. 3. We distinguished between the 6 native Dutch speakers and the 9 foreigners and the averages and standard deviations for both groups are given side by side. For each native Dutch the standard for a vowel was determined as the average of the vowel spectra spoken by the other five. For the foreigners all six native Dutch made up the standard. The averages varied from 9 to 15 dB for the native Dutch and from 14 to 29 dB for the foreigners, while for each individual vowel the
The average for the foreigners was higher. If all 16 filters contribute equally to the distances, the average contribution per filter is 2.8 dB for the native Dutch and 4.2 dB for the foreigners.

These results agree with our perception, as the pronunciation of the foreigners was, in general, of a worse quality. So far, however, the conclusion is only based on informal listening. A thorough perceptual verification must follow. Therefore an experiment is carried out in which sixty advanced students of an institute for speech therapists have to judge the quality of these vowels. Hereby the method of paired comparisons is used. Data from this experiment will be compared with the spectral distances in order to investigate the feasibilities of the analysis system.

Distance measures must yield optimal agreement between perceptual and physical data. Weighting functions of the 'Klatt-metric', which gave no better results than the Euclidian distance up till now, have to be optimized. Assmann and Summerfield [1] suggested to use the second derivative in order to give attention not only to spectral peaks but also to spectral 'shoulders'. Also a speech-independent weighting of the different filterbands may improve the results.

ACKNOWLEDGEMENT

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands organization for research, NWO.

REFERENCES

[1] ASSMANN, P.F. & SUMMERFIELD, Q. (1989) 'Modeling the perception of concurrent vowels: vowels with the same fundamental frequency' J.A.S.A. 85 pg 327-338


